

# DTW-MIC Coexpression Networks from Time-Course Data

Samantha Riccadonna<sup>1</sup>, Giuseppe Jurman<sup>1</sup>, Roberto Visintainer<sup>1,2</sup>,  
Michele Filosi<sup>1,3</sup>, Cesare Furlanello<sup>1</sup>

<sup>1</sup> Fondazione Bruno Kessler

Trento, Italy

{riccadonna,jurman,visintainer,filosi,furlan}@fbk.eu

<sup>2</sup> DISI

University of Trento, Italy

<sup>3</sup> CIBIO

University of Trento, Italy

## Abstract

When modeling co-expression networks from high-throughput time course data, the Pearson correlation function is reasonably effective. However, this approach is limited since it cannot capture non-linear interactions and time shifts between the signals. Here we propose to overcome these two issues by employing a novel similarity function, DTWMIC, combining a measure taking care of functional interactions of signals (MIC) and a measure identifying horizontal displacements (DTW). By using a network comparison metric to quantify differences, we show the effectiveness of the DTWMIC approach on both synthetic and transcriptomic datasets.

## 1 Introduction

Inferring a biological graph (*e.g.*, a Gene Regulatory Network) from high-throughput longitudinal measurements of its nodes is nowadays one of the major challenges in computational biology, and several are the proposed solutions to this still unanswered question [De Smet and Marchal, 2010; Marbach *et al.*, 2010]. Although the presence of many indirect interactions set the problem in a strongly non-linear domain, simple basic approaches such as the coexpression networks via correlation measures proved to work reasonably well [Allen *et al.*, 2012]. However, Pearson's correlation lacks sensitivity in case of non-linear relations between signals and when signals are mutually delayed, thus the reliability of a coexpression network would benefit from being built by a measure taking care of these issues. A first example in this direction can be found in [ElBakry *et al.*, 2010], where the authors envisioned an *ad-hoc* technique to cope with the time lag.

Here we follow a different approach introducing a novel similarity measure DTWMIC, whose components are the Dynamic Time Warping (DTW, [Sakoe and Chiba, 1978]) and the Maximal Information Coefficient (MIC, [Reshef *et al.*, 2011]). While DTW can take care of the time-shifts, the MIC metric can detect functional non-linear interactions between the signals. Moreover, the presence of MIC alleviates the limitations

affecting DTW when signals, although related, present different levels of expression. To support the claim of better reconstruction performances (in terms of smaller HIM distances from the gold standard), we use three synthetic datasets generated by GeneNetWeaver (GNW, [Schaffter *et al.*, 2011]) following the guidelines of the DREAM Challenge [Prill *et al.*, 2010], and a transcriptomic dataset on the expression response of human T cells to PMA and ionomycin treatment [Rangel *et al.*, 2004].

## 2 Methods

### 2.1 Coexpression Networks

Correlation methods like WGCNA [Zhang and Horvath, 2005] represent the most direct approach to the exploration of the gene co-expression network. The adjacency matrix is defined by computing a correlation function (absolute Pearson for WGCNA) between all pairs of gene signals  $G_i, G_j$ , soft thresholded (for instance by a power function) to determine the biological meaningfulness of the connections:

$$a_{ij} = f(G_i, G_j)^\beta.$$

These co-expression-based methods have been used in several studies and have shown their usefulness in interpreting biological results and identifying important gene modules. Therefore, we will use WGCNA as a reference method for the correlation-based approach.

To take care of non-linear interactions and time shifts in the gene signals, we propose to employ the DTWMIC function:

$$\begin{aligned} a_{ij} &= \text{DTWMIC}(G_i, G_j) \\ &= \left( \frac{1}{\sqrt{2}} \sqrt{\text{DTW}_s(G_i, G_j)^2 + \text{MIC}(G_i, G_j)^2} \right)^\beta. \end{aligned}$$

In all experiments we set  $\beta = 6$  (as in the WGCNA approach) for both Pearson and the DTW-MIC measures. Comparison with gold standard network is quantitatively assessed in terms of HIM network distance. Details on DTW, MIC and HIM are provided hereafter.

### 2.2 Maximal Information Coefficient

The Maximal Information Coefficient (MIC) measure is a component of the Maximal Information-based Nonparametric Exploration (MINE) family of statistics, introduced

in [Reshef *et al.*, 2011; Speed, 2011], for the exploration of two-variable relationships in multidimensional data sets. having the generality and equitability property. The MIC value is obtained by building several grids at different resolutions on the scatterplot of the two variables and computing the largest possible mutual information achievable by any grid applied to the data, and then normalizing to the  $[0, 1]$  range.

The two distinctive features of MIC are generality, *i.e.*, the ability of capturing variable relationships of different nature and equitability, that is the property of penalizing similar levels of noise in the same way, regardless of the nature of the relation between the variables. MIC can be computed in R by using the *minerva* package [Albanese *et al.*, 2012].

### 2.3 Dynamic Time Warping

The Dynamic Time Warping (DTW) [Keogh and Pazzani, 1998; Keogh and Pazzani, 2000] is a measure of distance between two sequences considering occurring time shifts between the series. Thus it proves more suitable than the Euclidean metric in curve comparison because it takes into account the shapes of the curves instead of just evaluating the pointwise distance of the vectors.

The DTW algorithm also finds an optimal match between the two given series by non-linearly warping them in the time dimension to determine a measure of their dissimilarity, stretching (or compressing) the time axis. As a comprehensive reference, the reader is referred to [Gusfield, 1997].

DTW elective application is the comparison of different speech patterns in automatic speech recognition [Sakoe and Chiba, 1978], but it has been also used in functional genomics [Aach and Church, 2001; Furlanello *et al.*, 2006]. To obtain a similarity measure, we use the function  $DTW_s = 1/(1 + DTW_d)$ , where  $DTW_d$  is the normalized distance between two series, as computed in the R package *dtw* [Giorgino, 2009].

### 2.4 Hamming Ipsen Mikhailov Distance

The HIM distance for network comparison is defined as the product metric of the Hamming distance  $H$  [Tun *et al.*, 2006; Dougherty, 2010] and the Ipsen-Mikhailov distance  $IM$  [Ipsen and Mikhailov, 2002], normalized by the factor  $\sqrt{2}$  to set its upper bound to 1:

$$HIM(N_1, N_2) = \frac{1}{\sqrt{2}} \sqrt{H(N_1, N_2)^2 + IM(N_1, N_2)^2},$$

for  $N_1, N_2$  two undirected (possibly weighted) networks. The drawback of edit distances (such as  $H$ ) is locality, as it focuses only on the network parts that are different in terms of presence or absence of matching links [Jurman *et al.*, 2011].

Spectral distances like  $IM$  are global, since they take into account the whole graph structure, but they cannot distinguish isomorphic or isospectral graphs, which can correspond to quite different conditions within the biological context.

Thus the HIM distance is a possible solution for both issues: details about HIM and its two components  $H$  and  $IM$  are given in [Jurman *et al.*, 2012].

## 3 Results

### 3.1 GeneNetWeaver data

GNW allows the construction of DREAM-like networks and corresponding simulated expressions (steady states and time course) generated as subnetworks of the *E. coli* [Gama-Castro *et al.*, 2008] or the Yeast [Balaji *et al.*, 2006] transcriptional regulatory network, with the possibility of choosing a number of parameters both for the subnetwork and the data structure. In particular, we generated three synthetic networks Yeast<sub>20</sub>, Ecoli<sub>20</sub>, Ecoli<sub>50</sub>, where the subscript indicates the number of nodes of the the net, extracted randomly from the whole set of nodes. In each case, half of the selected nodes are regulators.

For each datasets, we generated 10 replicates of longitudinal datasets, created by a dynamic model mixing ordinary and stochastic differential equations, with 41 time points equally spaced between time 0 and time 1000 and affected by 0.5% (Yeast) or 1% noise (*E. coli*). In each dataset, the first half of the time series shows the response of the network to a perturbation (at  $t=0$  is the wild-type steady-state), then the perturbation is removed and the second half of the time series shows how the gene expression levels go back from the perturbed to the wild-type state. (this is the DREAM4 setup for the expression time course in GNW).

For each experiment, we built the corresponding Pearson and DTWMIC coexpression networks, and we computed the HIM distance of both the inferred networks from the GNW gold standard. The results are reported in Table 1 and they show that the DTWMIC networks are consistently closer to the gold standard than the corresponding Pearson WGCNA graph.

For the Yeast<sub>20</sub> dataset, we further generated 4 time course datasets of the same length as above, but with a dual gene knockout. Also in this configuration, the DTWMIC inferred networks were closer to the gold standard: in the 4 experiments, their HIM distances were 0.23, 0.21, 0.24 and 0.21, while for the Pearson correlation networks the corresponding values were 0.30, 0.27, 0.31 and 0.47 respectively.

# exp	Yeast <sub>20</sub>		Ecoli <sub>20</sub>		Ecoli <sub>50</sub>	
	P	D	P	D	P	D
1	0.57	0.23	0.37	0.20	0.22	0.22
2	0.41	0.21	0.41	0.19	0.31	0.20
3	0.39	0.20	0.37	0.19	0.23	0.23
4	0.25	0.25	0.23	0.36	0.27	0.21
5	0.56	0.24	0.41	0.19	0.35	0.19
6	0.35	0.19	0.53	0.20	0.40	0.20
7	0.56	0.23	0.40	0.19	0.26	0.21
8	0.42	0.22	0.52	0.20	0.29	0.21
9	0.49	0.22	0.42	0.20	0.25	0.21
10	0.53	0.22	0.22	0.28	0.35	0.20

Table 1: HIM distances of the DTWMIC (D) and the WGCNA Pearson (P) networks for all experiments on the GNW datasets Yeast<sub>20</sub>, Ecoli<sub>20</sub>, Ecoli<sub>50</sub>

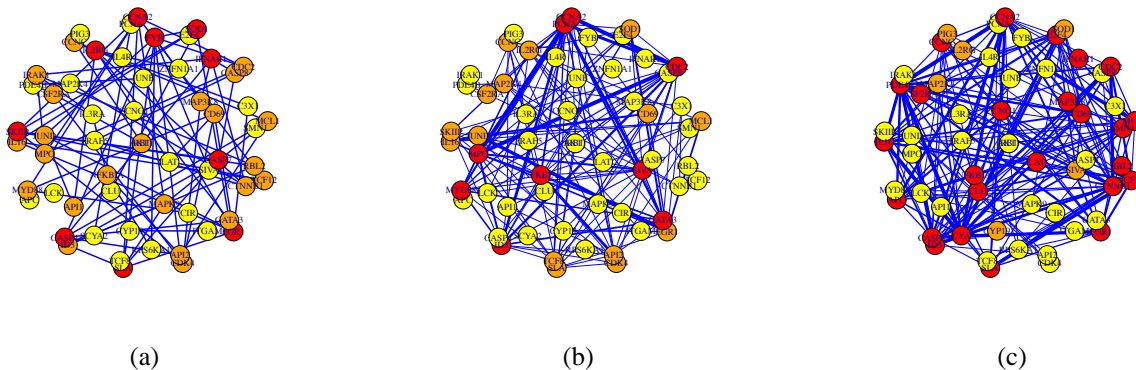


Figure 1: The T-cell network: gold standard (a), Pearson correlation network (b) and DTWMIC coexpression network (c). The gold standard has unweighted links, and the node color is red for degree  $> 10$ , orange for  $5 < \text{degree} < 10$  and yellow for node degree  $< 5$ . For the Pearson correlation network we show only the links whose weight is larger than 0.05; nodes are red when degree  $> 20$ , orange for  $10 < \text{degree} < 20$  and yellow for degree  $< 10$ . For the DTWMIC network we show only the links whose weight is larger than 0.15; node colors are the same as in the Pearson net.

### 3.2 T-cell data

In the paper [Rangel *et al.*, 2004], the authors investigated the response of a human T-cell line (Jirkat) to a treatment with PMA and ionomycin by measuring the expression of 58 genes across 10 time points (0, 2, 4, 6, 8, 18, 24, 32, 48, and 72 hours after treatment) with 34 replicates (data available in the R package *longitudinal*). Opgen-Rhein and Strimmer in [Opgen-Rhein and Strimmer, 2006b; Opgen-Rhein and Strimmer, 2006a] constructed the corresponding network by shrinkage estimation of the (partial) dynamical correlation, which we consider here as the ground truth network, displayed in Figure 1(a).

We then built, starting from the same dataset, the correlation networks inferred by Pearson and DTWMIC, plotted respectively in Figure 1, panels (b) and (c) respectively. Again, the DTWMIC network results closer to the gold standard than the Pearson net, with a HIM value of 0.164 versus 0.214.

Moreover, it is worthwhile noting what happens considering separately the two components of the HIM distance in this case: while the Ipsen Mikhailov distance is still smaller for the DTWMIC network (IM=0.203 vs. IM=0.296), the Hamming distance is larger (H=0.112 vs. H=0.06). This yields that there is a smaller number of links changing between the Pearson coexpression network and the gold standard, but these changing links induce a strongly different structure between the two graphs.

## 4 Conclusion

We introduced here DTWMIC, a novel measure for inferring coexpression networks from longitudinal data as an alternative to the absolute Pearson correlation used in the WGCNA approach. Due to the nature of its components

Dynamic Time Warping and Maximal Information Coefficient, the DTWMIC similarity can overcome the well known limitations of the Pearson correlation when dealing with horizontally displaced signals and indirect interactions. Experiments on biologically inspired synthetic data and gene expression time course show the higher precision in the network inference achieved by DTWMIC with respect to the Pearson correlation in different conditions.

### Acknowledgments

The authors acknowledge funding by the EU FP7 project HiperDART.

### References

- [Aach and Church, 2001] J. Aach and G. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.
- [Albanese *et al.*, 2012] D. Albanese, M. Filosi, R. Vissintainer, S. Riccadonna, G. Jurman, and C. Furlanello. cmime, minerva & minepy: a C engine for the MINE suite and its R and Python wrappers. arXiv:1208.4271 [stat.ML], 2012.
- [Allen *et al.*, 2012] J.D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao. Comparing Statistical Methods for Constructing Large Scale Gene Networks. *PLoS ONE*, 7(1):e29348, 2012.
- [Balaji *et al.*, 2006] S. Balaji, M.M. Babu, L.M. Iyer, N.M. Luscombe, and L. Aravind. Comprehensive Analysis of Combinatorial Regulation using the Transcriptional Regulatory Network of Yeast. *Journal of Molecular Biology*, 360(1):213–227, 2006.

- [De Smet and Marchal, 2010] R. De Smet and K. Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717–729, 2010.
- [Dougherty, 2010] E.R. Dougherty. Validation of gene regulatory networks: scientific and inferential. *Briefings in Bioinformatics*, 12(3):245–252, 2010.
- [ElBakry *et al.*, 2010] O. ElBakry, M.O. Ahmad, and M.N.S. Swamy. Inference of gene regulatory networks from time-series microarray data. In *Proc. 8th IEEE NEWCAS Conference*, pages 141–144. IEEE, 2010.
- [Furlanello *et al.*, 2006] C. Furlanello, S. Merler, and G. Jurman. Combining feature selection and DTW for time-varying functional genomics. *IEEE Transactions on Signal Processing*, 54(6):2436–2443, 2006.
- [Gama-Castro *et al.*, 2008] S. Gama-Castro, V. Jimnez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M.I. Pealozza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muiz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Ros, E. Morett, E. Merino, A.M. Huerta, L. Trevio-Quintanilla, and J. Collado-Vides. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research*, 36(suppl 1):D120–D124, 2008.
- [Giorgino, 2009] T. Giorgino. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7):1–24, 2009.
- [Gusfield, 1997] D. Gusfield. *Algorithms on strings, trees and sequences*. Cambridge University Press, 1997.
- [Ipsen and Mikhailov, 2002] M. Ipsen and A.S. Mikhailov. Evolutionary reconstruction of networks. *Phys. Rev. E*, 66(4):046109, 2002.
- [Jurman *et al.*, 2011] G. Jurman, R. Visintainer, and C. Furlanello. An introduction to spectral distances in networks. *Frontiers in Artificial Intelligence and Applications*, 226:227–234, 2011.
- [Jurman *et al.*, 2012] G. Jurman, R. Visintainer, S. Riccadonna, M. Filosi, and C. Furlanello. A global distance for network comparison. arXiv:1201.2931 [math.CO], 2012.
- [Keogh and Pazzani, 1998] E. Keogh and M. Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In AAAI Press, editor, *Proc. KDD '98*, pages 239–241, 1998.
- [Keogh and Pazzani, 2000] E. Keogh and M. Pazzani. Scaling up dynamic time warping for datamining applications. In AAAI Press, editor, *Proc. KDD '00*, pages 285–289, 2000.
- [Marbach *et al.*, 2010] D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, 107(14):6286–6291, 2010.
- [Opge-Rhein and Strimmer, 2006a] R. Opge-Rhein and K. Strimmer. Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65, 2006.
- [Opge-Rhein and Strimmer, 2006b] R. Opge-Rhein and K. Strimmer. Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. In P. Ruusuvuori, T. Manninen, H. Huttenen, M.-L. Linne, and O. Yli-Harja, editors, *Proc. 4th International WCSB 2006*, pages 73–76, 2006.
- [Prill *et al.*, 2010] R.J. Prill, D. Marbach, J. Saez-Rodriguez, P.K. Sorger, L.G. Alexopoulos, X. Xue, N.D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. *PLoS ONE*, 5(2):e9202, 02 2010.
- [Rangel *et al.*, 2004] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotharan, A. Gaiba, D.L. Wild, and F. Falciani. Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, 2004.
- [Reshef *et al.*, 2011] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. Detecting novel associations in large datasets. *Science*, 6062(334):1518–1524, 2011.
- [Sakoe and Chiba, 1978] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.
- [Schaffter *et al.*, 2011] T. Schaffter, D. Marbach, and D. Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.
- [Speed, 2011] T. Speed. A Correlation for the 21st Century. *Science*, 6062(334):1502–1503, 2011.
- [Tun *et al.*, 2006] K. Tun, P. Dhar, M. Palumbo, and A. Giuliani. Metabolic pathways variability and sequence/networks comparisons. *BMC Bioinformatics*, 7(1):24, 2006.
- [Zhang and Horvath, 2005] B. Zhang and S. Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(Issue 1):Article 17, 2005.